AN INTELLIGENT SYSTEM FOR ENHANCING MEDICAL BIG DATA SECURITY FRAMEWORK USING MACHINE LEARNING

A SYNOPSIS

Submitted by

RAJESHKUMAR K
(Reg. No. 9200104007)

in partial fulfillment for the award of the degree of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

KALASALINGAM ACADEMY OF RESEARCH AND EDUCATION

(DEEMED TO BE UNIVERSITY)

ANAND NAGAR, KRISHNANKOIL – 626 126

TAMIL NADU, INDIA

JANUARY 2024

1. INTRODUCTION

Big data (BD) has meaningfully transformed how industries handle, scrutinize, and employ data in each area. Healthcare is one such industry in which BD plays a vital role in changing things. Big data in healthcare helps enhance patient outcomes, identify disease spread, gather essential data, prevent patients from serious health issues, diminish healthcare delivery costs, and improve patient's quality of life [1]. With the enormous development of the healthcare industry, the Big data period commonly requires the increasing use of cloud systems to host information and permit easy access to it over the web. Combining cloud computing with Big dat networks helps to encompass flexibility, scalability, resource virtualization, and parallel processing [2, 3]. A large amount of data is being shared and uploaded to the cloud by medical users, and it is challenging to manage these large amounts of Big dat in the cloud effectively. *Apache Hadoop* is an effective software mechanism that helps process large amounts of Big dat by employing the Hadoop distributed file system (HDFS) and the MapReduce (MR) program. The HDFS is responsible for partitioning and administrating the data, whereas the MR is a progressive model that controls the processing part. It is challenging to manage these distributed computing platforms to enhance the mining effects of standard data mining models as well as promptly offer data services on the complex Big dat platform [4 and 5]. Clustering plays a significant role in effectively managing the vast medical data on the Hadoop platform [6].

The widely used clustering models by the researchers for Hadoop clustering are K-means (KM), fuzzy c-means (FCM), k-medoids, density-based clustering, COBWEB, and grid-based CLIQUE [7]. Like Big dat management, Big dat security is a significant challenge to protecting healthcare organizations from intruders. Because more information is stored in a cloud environment using the Hadoop structure, its flexibility makes it more adaptable and raises security constraints [8]. When using these software platforms for processing a large amount of data in the cloud, the possibility of attackers getting access to modify that stored data increases. So, it is necessary to offer a secure mechanism to protect medical data from intruders because the medical records uploaded to the cloud include patients' details like medical histories, treatments, medical claims, etc. [9]. The healthcare industry manages these medical data for several reasons, such as research, billing, and storage. So, security is vital for HC administrations to safeguard patient's information.

The researchers widely used encryption techniques to secure patients' medical information. The encoding performed on the raw medical data contains an unreadable data format the human does not understand. If any of them want to get it in an original form, they require a key from the authorized individual. Otherwise, they will not be able to access it. However, more than encryption is needed for Big data security in the cloud. Providing secure access to the data allows only organized persons in the network to protect the data further. The analysis clearly shows that the challenges we must focus on in an enabled cloud environment are data management, security, and access. Our study aims to satisfy these challenges by proposing a cluster-based optimized MR model to manage the Big data and hashing and asymmetric-based encryption approaches effectively to offer a highly secure environment for the medical BD on the cloud.

2. LITERATURE SURVEY

Here, the survey of recent Big data management and security methods using clustering, encryption and access control mechanisms is discussed. The Big dat classification schemes developed for big medical data using machine and deep learning models are also discussed. The survey addresses the limitations caused by the existing research frameworks and provides solutions to overcome those problems efficiently.

Fouad Awad and Murtadha Hamad [10] presented a Big data clustering approach using neural-processor-based KM that considered the advantages of mobile devices' machine learning processors. The outcomes illustrated that the model improved the clustering speed a and performance using neural processors compared to the traditional processors. Josephine Usha Lawrance and Jesu Vedha Nayahi Jesudhasan [11] suggested a parallel clustering-based anonymization algorithm (PCAA) for Big data clustering using MR paradigm. The system was scalable and attains a trade-off between utility and privacy. The large quantity of data is processed parallelly using the MR model and the model attained the better performance regarding f-score and accuracy. Dandug LS and Vani KS [12] presented a weighted KM model for clustering the Big data based on MR. The system used Aquila optimization model to optimally selects the cluster centroids of the clustering technique. Ramdas Vankdothu *et al*. [13] presented an entropy and sigmoid based KM clustering approach for handling Big dat in HDFS. The model used adaptive grey wolf optimization (AGWO) algorithm to choose a suitable machine for distributing the data blocks into it.

Shanmugapriya and Kavitha [14] presented a medical Big dat security system using encryption with access control mechanism. The Big data clustering was performed using

optimal fuzzy c means (OFCM) and the clustered data was encrypted using triple data encryption standard (TDES) to encrypt and store the data in the cloud. Ila Chandrakar, and Vishwanath Hulipalled [15] presented a two-level privacy system using pseudonymization and homomorphic encryption. The system reduced the execution time by 50% and enhance privacy by 10% for BD security and mining. Siti Dhalila Mohd Satar *et al.* [16] offered fine grained access control and data integrity in Big data using ciphertext policy Attribute based encryption (CP-ABE). Kashish Shakil *et al.* [17] presented a biometric authentication model for securing the BD environment. The system used resilient back propagation neural network to perform training of the biometric signature samples on HDFS. The method attains the sensitivity and specificity of 0.98 and 0.95, which was better than previous research models. Sunday Oyinlola Ogundoyin and Ismaila Adeniyi Kamil [18] presented a lightweight privacy-preserving authentication and fine-grained access control scheme for smart health called PAASH.

Yongsheng Zhou and Majid Ghani Varzaneh [19] suggested a secure BD clustering model in cloud platform using locality-sensitive hashing approach. Pravin Game et al. [20] suggested a BD classification system using support vector machine (SVM). The model used principal component analysis for reducing the dimensions of the dataset and divergence based grey wolf optimization algorithm was utilized for optimization of the classifier. Jian Ping Li *et al*. [21] presented an ensemble of ML models such as logistic regression, SVM, k-nearest neighbour, and artificial neural network for Big data classification. Comparing all SVM provides better accuracy than other ML models.

Problem Statement

- The existing clustering approaches selects cluster centroids in a random manner that affects the clustering performance regarding convergence and local optima. Additionally, data deduplication was not focused. The patients' medical data contains more redundant and duplicate data that needs to be removed to attain higher prediction outcomes.
- The encryption models called TDES and homomorphic encryption for BD security takes higher computational time and limited estimation processes than similar encryption techniques that impact the algorithm's efficiency. It also requires more computing power and memory storage. Also, due to the centralized nature of

CP-ABE, it is moderately complex, and data alteration is possible without the owner's permission.

- The researchers mostly use single sources for access control, which are simple and user-friendly. However, the likelihood of attacks is high in these single source-based systems because the intruders can easily guess the details of the users and alter important information. So, using a multi-factor rather than a single-factor model is better.
- The BD classification models provides satisfying outcomes; however, manual feature learning in machine learning-based models is time-consuming, and they generate lower outcomes when large amounts of data are used. Deep learning is better than machine learning models for automated feature learning and attains higher accuracy in BD classification.

3. RESEARCH OBJECTIVES

- To ensure the secrecy of the users in medical Big data by employing a novel three-factor authentication (3FA) and levy flight and scoring mutation-based chimp optimization (LFCOA) technique for optimal BD management in HDFS.
- To propose a secure hash algorithm XOR with ECC (SHAXECC) to protect medical big data from unauthorized users.
- To develop a cluster-based optimal MR model for Big data management in HDFS and Whirlpool Hashing-based Enhanced Rivest Shamir Adelman (WHERSA) for providing secure medical BD storage in the cloud.

4. RESEARCH METHODOLOGY

Contribution 1: 3FA and LFCOA based Optimal Clustering for Maintaining User Privacy in Cloud

BD is a huge as well as miscellaneous, which is increasing exponentially. BD in healthcare offers several benefits to the fields, such as enhancing patient outcomes, attaining valuable insights, evading preventable diseases, dropping healthcare costs, predicting widespread outbreaks, and improving the quality of patient's life. Cloud computing is an adequate environment for processing, analysing and storing big medical data. Simultaneously, the BD stored in the cloud is prone to several security threads, so it is necessary to prevent healthcare records from unauthorized access. This study proposes a secure BD management system in

the cloud using novel authentication mechanism. The study includes three phases: user authentication, data deduplication, and data management. Figure.1 illustrates the proposed work's flow diagram. Initially, the authentication of the BD users is done using a novel 3FA mechanism to preserve privacy in access control of the data in the cloud. Then, data deduplication of the user's data is done using a hashing mechanism that prevents the redundant data of the users from being continuously stored in the cloud. Finally, data management using optimal clustering was done to manage the medical BD in the MR framework efficiently.

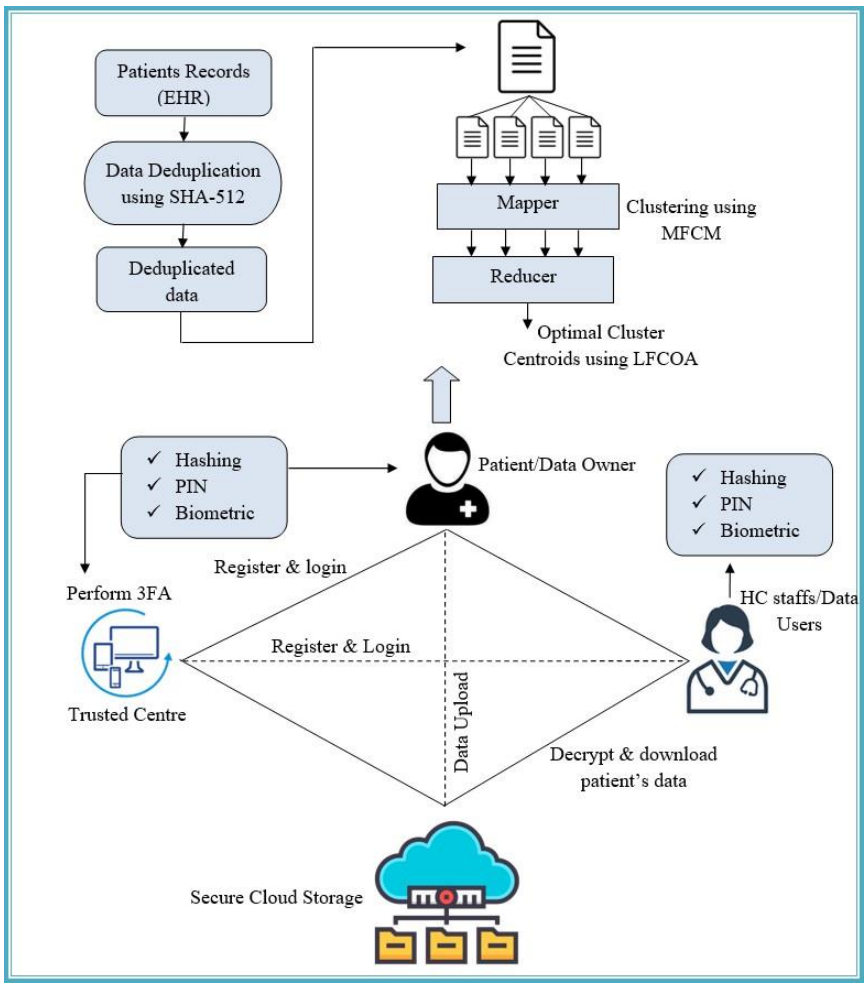

Figure 1: Proposed system's structural design

User authentication

User authentication helps the medical Big data environment give only authorised individuals access, thereby increasing the protection of patients' healthcare records. The authentication of the healthcare users in the current study is carried out by performing three steps: registration, login, and verification, which are described as follows.

(a) User registration

If the medical users say patients and healthcare specialists require data access, they are initially registered with the cloud's trusted center (TC) to perform verification. The user initially offers the details like user name (U_{xna}), user id (U_{xid}), age (A_{xa}), mobile number (M_{xnum}), password (P_{xpw}), biometric (fingerprint) (B_{xau}), and timestamp (T_{xts}) to the TC for performing registration. Then, the hash value is generated by the TC by employing secure hashing algorithm-512 (SHA-512) that combines U_{xid}, P_{xpw}, as well as salt value (S_{xval}) that ranges between 0 and 1. The salt value considered during registration is declared as the user's login secret key to perform login. The rationale for preferring the SHA-512 model is that it is not "broken" compared to other traditional hash algorithms and has no known security-compromising weaknesses. The input of the hashing model is chunks of 1024-bit data, and the output will be 512-bit cipher data. The hash value H_{xval} produced using SHA-512 hash function h_{xhf} is shown in eqn. (1).

$$H_{xval} = h_{xhf}(U_{xid} \,||\, P_{xpw} \,||\, S_{xval}) \tag{1}$$

The above hash data and the random salt value (login secret key) is stored in cloud for performing authorization of the BD users.

(b) Login and verification

After registration, the system authenticates the Big data users by conducting three levels of authentication procedures.

- Firstly, the system authorizes the users based on the hash data. The system asks the users to enter U_{xid}, $P_{xpw}$and S_{xval}. The hash value for the entered details is created using SHA-512 hashing, and the created one is compared with the one generated and stored in the cloud during user registration. If both values match values match, the system allows the users to perform the second authentication phase.
- Secondly, a one-time password is sent to the mobile number of the registered user, and the user wants to provide the received OTP to the system for verification. If both are matched, this authentication stage is successful, and they are allowed to do the final stage.

- Finally, the user's biometrics records are used to check their authority. If the entered biometric sample matches the database one, they are considered a trusted user and allow medical data access in the cloud.

In each phase of authentication, only the users authorized by the system are allowed to perform the next level of authentication. Otherwise, they are considered unauthorized or forged, and access is denied to that user.

Data deduplication

The system allows the users to access the medical Big data stored in the cloud once authorized using our 3FA model. Medical organizations upload enormous amounts of medical data to the cloud for several purposes. It takes work to deal with this large amount of data to offer effective services to the users promptly. So, these enormous amounts of data can be handled using the HDFS in the cloud by providing fault-tolerant, scalable, and cost-efficient storage for BD. In our study, the deduplication concept is implemented to diminish the HDFS' storage space in the cloud by avoiding redundant medical data of the users being stored in the cloud. The user files are split into several chunks, and the hash value of each chunk is calculated and contrasted using the hash value. If any files containing the same hash value are regarded as duplicate data, they are removed. Otherwise, the files are uploaded successfully.

Data management

Once the data is deduplicated, it is processed using our optimal MR technique to manage the data effectively in the cloud. The proposed MR system consists of two stages to manage the medical BD optimally: mapper and reducer. The general structure of MR-based data management for healthcare data is given in Figure.2.

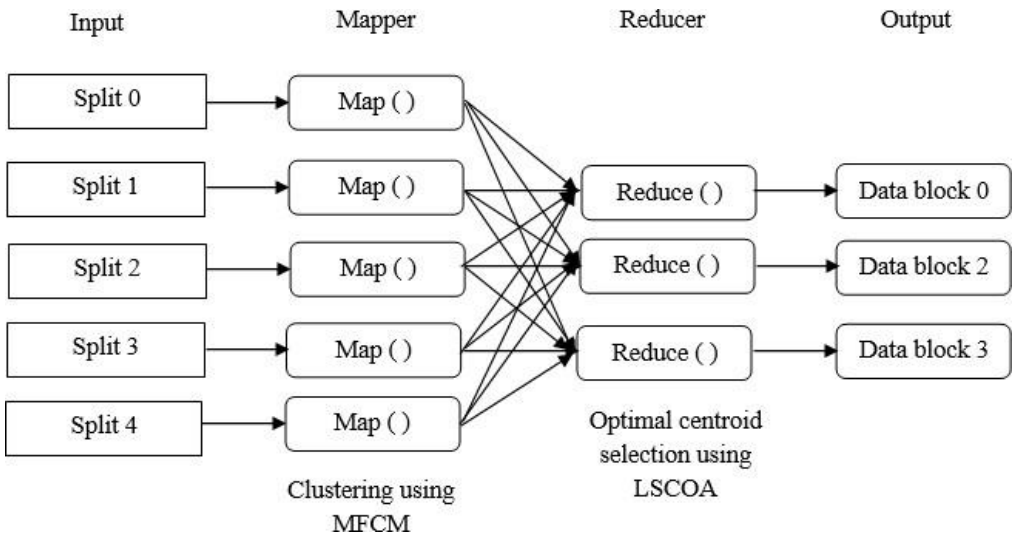

Figure 2: Proposed MR framework

(a) Clustering in mapper

The mapper phase mainly performs clustering of the Big data that computes the distance between the data points in the given data and groups similar data points into a cluster of groups. The study uses a modified FCM (MFCM) to cluster the medical data of the healthcare organizations based on the membership function. The conventional FCM clustering technique suffers from the following drawbacks: it clusters the data points within the group, has higher error rates, and needs more iteration to get effective clusters. Our study develops the MFCM technique that overcomes the problem of clustering the data points within the group because there is a possibility of having similar data points in more than one cluster group, which helps to converge faster and lower the error rate.

(b) Optimization in reducer

This stage receives the output from the mapper stage and optimizes to decide on a final set of clusters to be managed in a cloud server. The study introduces the levy flight and scoring mutation-based Chimp optimization algorithm (LSCOA) for performing optimization in a reducer that selects the optimal cluster centres for the cluster groups retrieved from the mapper stage. These optimization-based clusters formed in the MR framework led to higher performance in data management by considering the traits of their workloads. The optimization model presented in our study is only applied with improvement. Since the conventional Chimp optimization algorithm (COA) suffers from the local optimal problems

and a slower convergence rate. So, we are adopting levy flight (LF) and scoring mutation (SM) functions to enhance the performance of the COA further. The utilization of LF in COA makes the searching process more global and safeguards the COA from being trapped into local optimal solutions.

Meanwhile, the SM in COA helps improve the convergence speed and selection probability. So, the name LSCOA is given because of the incorporations of LF and SM in the algorithm's traditional structure. Figure.3 demonstrates the pseudocode of the proposed LSCOA for optimal centroid selection.

Input: Cluster set by the mapper

Output: Optimal cluster set

Begin

Initialize the population size and maximum number iterations
Compute each chimp's fitness using DB-index
Select attacker, barrier, chaser, and driver
Sort the population by scoring mutation
While $n < \max num\ of\ iter$
For each chimp
If random number $\kappa < 0.5$
Update the current chimp's position based on the position of attacker, barrier, chaser, and driver
Else if
Update the current chimp's position using levy flight and chaos model
End if
End for
Update the dynamic coefficients
Compute fitness of each chimp
$n = n + 1$
End while
Return $M_{attacker}$

End

Figure 3: Pseudocode of LSCOA

Evaluation

Here, the proposed research system is evaluated by contrasting their outcomes with the prevailing research techniques. The system uses PYTHON as an implementation platform with an Intel i5/core i7CPU processor, Windows 7 OS, 4GB RAM and 3.20 GHz speed. The

research uses an openly available heart disease dataset obtained via Kaggle to train and test the model. It encompasses 76 traits like sex, age, type of chest pain, heart rate, class label etc.

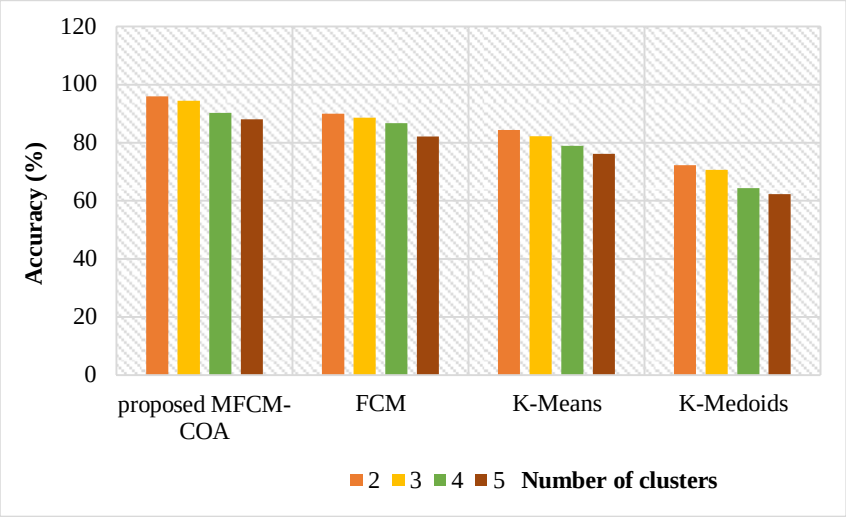

Figure 4: Accuracy analysis

Figure. 4 shows the accuracy obtained by the proposed (MFCM and LSCOA-based BD management named MFCM-COA) along with prevailing models for BD management. The MFCM-COA reached a maximum accuracy of 96% for cluster size 2. However, the existing KM, FCM, as well as K-medoids provide an accuracy of 84.43%, 90.26%, and 72.26%, lower than MFCM-COA. Likewise, our proposed optimal MR-based clustering technique attains higher accuracy (in every cluster size) for performing clustering tasks in HDFS.

Contribution 2: Secure Cloud Storage using SHAXECC

It is essential to encode the data in the HDFS cloud to protect it from third parties and intruders once the data management of Big data is done. The research work introduces the SHAXECC mechanism for encoding the medical Big data. Elliptic curve cryptography (ECC) is an asymmetric encoding technique based on the theory of elliptic curves. The technique offers a faster, smaller, and more effective critical generation process for encoding and decoding. To further enhance the security of the ECC model, our study makes some improvements. Generally, public and private keys are employed in ECC for cryptographic operations. Our model uses a hash-based secret key (produced during the registration process) to enhance the system's security further. This key is used in both the encryption and decryption processes instead of only considering ECC's public and private keys. Also, the private key of the ECC is generally created based on a random number. Here, we are

performing the XOR operation between the random number and secret key to obtain the final private key used for further cryptographic processes. So, the SHA and XOR-based key generation and encryption in ECC is called SHAXECC. The flowchart of the SHAXECC is shown in Figure.5.

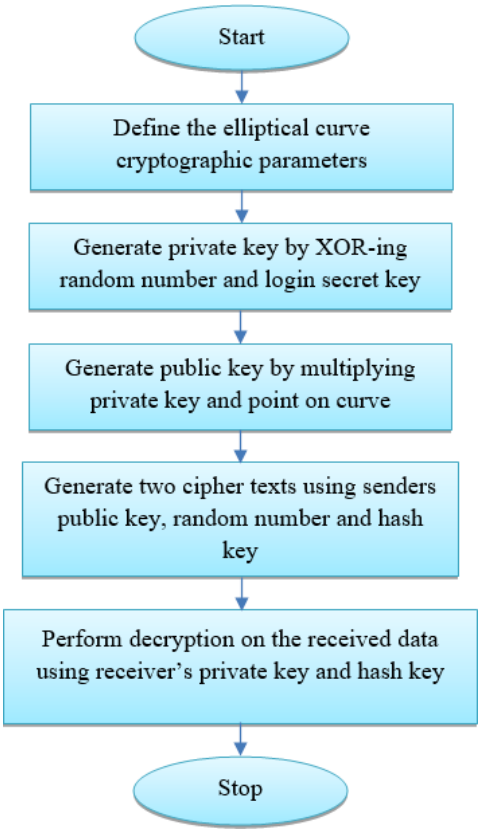
Start
Define the elliptical curve cryptographic parameters
Generate private key by XOR-ing random number and login secret key
Generate public key by multiplying private key and point on curve
Generate two cipher texts using senders public key, random number and hash key
Perform decryption on the received data using receiver's private key and hash key
Stop

Figure 5: Flowchart of SHAXECC

Evaluation

The proposed encryption technique called SHAXECC is compared with the existing techniques, say ECC, rivest shamir adleman (RSA), and advanced encryption standard (AES) regarding encryption and decryption time by varying several Big data files from 100 to 500 Mb, which is shown in Figure.6. The SHAXECC takes 53,643ms to encrypt the data and 38,114ms to decrypt the data when the file size is 100 Mb, which is minimal compared to the existing approaches.

(a) (b)

Figure 6: Encryption and decryption time analysis

Because the existing RSA, AES, and ECC take 58,567ms, 56,897ms, and 55,567ms to encrypt the data and 43,321ms, 41,123ms, and 40,897ms to decrypt the data for the same 100Mb data file. Similarly, our model also offers less time to perform encryption and decryption for the other files (200 to 500 Mb). Thus, these outcomes for data encryption show the effective cryptographic procedure of SHAXECC with minimal computation overhead.

Contribution 3: Cluster-based Optimal MR model for BD Management and WHERSA based Secure Medical Storage

This contribution presents another variant of the optimal MR system with a hashing-based encryption mechanism and deep learning-based BD classification system for secure BD management and healthcare monitoring in the cloud. The system uses the blockchain mechanism initially to authenticate the cloud medical users. Then, BD management is performed by employing kernelized KM (KKM) and an enhanced butterfly optimization algorithm (EBOA). Then, the system used WHERSA to secure the HDFS data's cloud storage. Finally, BD classification is performed by employing an optimal deep-learning network to monitor the patients' health conditions. Figure.7 portrays the structural framework of the current research.

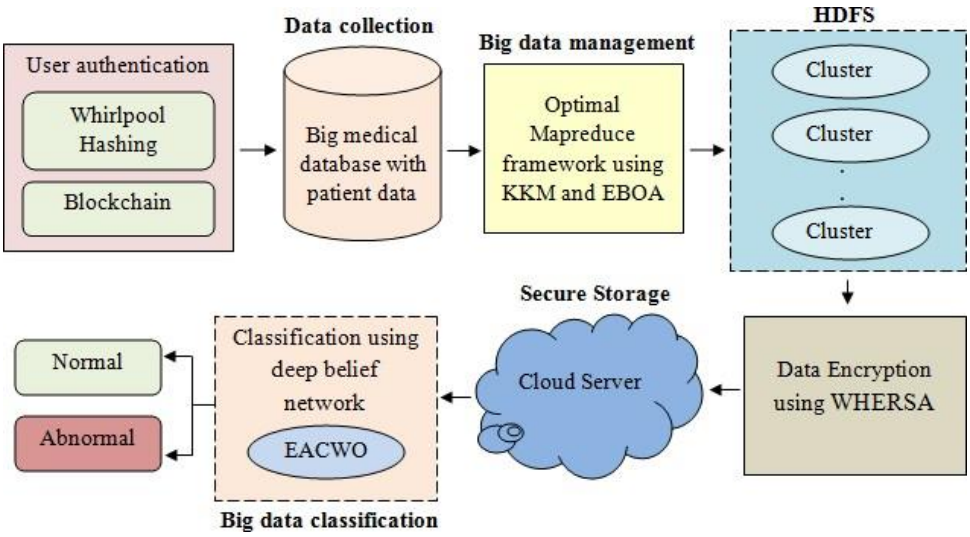

Figure 7: Structural framework of secure and optimal BD clustering using WHERSA

User authentication

The study aims to authenticate the medical Big data users using hashing and blockchain-based mechanisms. Initially, the system uses a whirlpool hashing technique to authorize the healthcare users in the BD environment. The system relies on TC, which performs registration, login, and verification.

- ***Registration***

The users are registered with the TC by providing their appropriate credentials. The TC then produces the hidden key by employing the user's ID, password as well as a random number and performs hashing on the generated hidden key using the Whirlpool hash (WH) function, which again concatenates the user ID and password with the hidden key to make the hash value stronger. The WH algorithm is more secure than other algorithms and returns the 512 bits of hash value for the given input. The working of the WH function is demonstrated in Figure. 8.

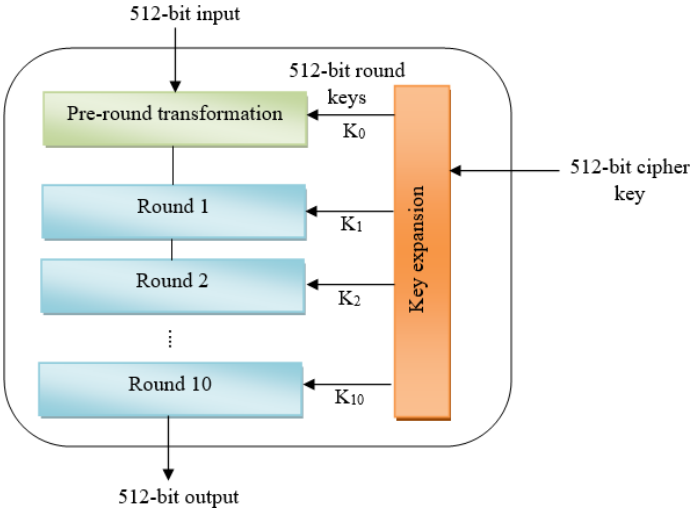

Figure 8: Working of WH function

Generally, the user credentials are kept in the cloud for authentication. However, there is a high risk of intruders getting these user credentials when storing the data in the cloud rawly. To avoid this, our model employs the blockchain system to protect the user credentials in the cloud. The blockchain system generates the hash values for each block of data it receives. It contains the chain of hash codes for the given data by maintaining the order of transaction information, hash code, and previous blocks' hash code value. If any intruder alters the data, it will inform the respective owner of the medical data that was stored in it. So, it is difficult to crack the data stored in the blockchain. Consequently, the system's security is enhanced using hashing and blockchain mechanisms.

- **Login and verification**

During login, the users enter the hidden key, user ID and password, and the TC produces the respective hash value of the given particulars using the WH function. The generated hashing value is compared with those generated and stored in the cloud at the time of user registration at TC. If both matches, the TC accepts the user for data access; otherwise, it rejects the user.

Big data management

After successful authentication, the users can get the permission to access their data in the cloud. The users' medical data comprises lakhs or millions of raw data in the repository. It is essential to preprocess the raw medical information for enhancing its quality and diminish the processing time for BD management and classification. The system performs missing value imputation, numerical conversion, along with normalization as the preprocessing steps.

The missing value imputation process eliminates the missing values by computing the mean values of the missing data column values. Then, the dataset values are changed into numerical values using a single hot encoder to fit into machine-readable form. Finally, min-max normalisation is performed to improve the training model's stability.

The pre-processed BD is fed into optimal MR design to manage the HDFS data effectively. In MR, the mapper phase performs the clustering of the BD using the KKM algorithm, and the reducer phase determines the optimal cluster centroids to choose the BD's final clusters using EBOA iteratively.

(a) Clustering in mapper

The proposed system uses KKM to cluster the data points into several sets that diminish the intra-cluster distance as well as maximize the inter-cluster distance. The traditional KM algorithm is faster than another clustering model. However, it estimates the similarity between the data points using Euclidean distance, which leads to the problem of the curse of dimensionality. So, our study uses the KKM clustering approach that is better than related clustering schemes by maintaining computational complexity.

(b) Centroid optimization in reducer

The clusters formed in the mapper stage are given to the reducer stage for choosing optimal centroids for the clustered data points that iteratively result in the final set of data points stored in the cloud. The study proposes EBOA for choosing optimal cluster centroids in the reducer. The butterfly optimization algorithm (BOA) mimics the foraging behaviour of butterflies for optimal engineering problems. Like the general optimization method, it tends to fall into local optimal solutions, has a moderate convergence rate at an earlier stage of the algorithm and is limited to global search capability at the later iteration period. To overcome those problems, chaotic mapping improves the initial population's accuracy. In addition, the model includes a dynamic switching probability in its structure that prevents the model from local optimal solutions and leads to faster convergence. These enhancements over traditional BOA are termed EBOA, and the EBOA's pseudocode is illustrated in Figure. 9.

Input: Cluster center by mapper

Output: Centroid optimization by reducer

Begin

Initialize the butterfly population using chaotic logistic mapping

$$\hat{m}_u^{\gamma+1} = \chi \hat{m}_0 (1 - \hat{m}_u), \quad 0 \le \hat{m}_0 \le 1$$

Evaluate the fitness of butterfly

While stopping criteria not met **do**

For each butterfly in population **do**

Calculate fragrance for butterfly based on stimulus intensity function

End for

Update the new position of the butterfly using dynamic switching probability

End while

Return optimal value

End

Figure 9: Pseudocode of EBOA

Secure data transfer

The information is securely stored in the cloud to safeguard it using the WHERSA technique. RSA is a public essential encryption technique that uses two types of keys, public as well as private keys, for encoding and decoding the data in the cloud. Using the public key, the sender encodes the user's information and using the private key, the recipient decodes the information. The key generation process in RSA is performed randomly using the prime numbers, which increases the chances of getting attacks on it. So, our model generates four prime numbers instead of two prime numbers and performs encryption and decryption based on that. Also, the hidden key (generated during the registration process) is used in the RSA cryptographic procedure to enhance the model's security further. These enhancements in RSA using WH are named WHERSA.

Big data classification

When the data is received from the hospital, it is necessary to check the patients' health condition using some advanced algorithms to take safety precautions on time. The study presents an optimal DBN for performing BD classification. The DBN is a deep learning approach that helps extract MR data's hierarchical structure. It uses hidden neurons to learn the data effectively and achieve a higher accuracy. To improve the performance of the DBN

further, our system uses the enhanced ant colony weight optimization (EACWO) technique for weight optimization. The model works by mimicking ant colonies' foraging behaviour and finding the optimal paths for a DBN. The conventional model of ant colony optimization has the problems of slower convergence and lower solution accuracy. So, the model is enhanced by including the Rosen Brock function and contraction factor that improves the algorithm's convergence by providing global search ability. This improvisation in traditional ant colony optimization is termed EACWO. The structure of DBN for accurate Big data classification is shown in Figure. 10.

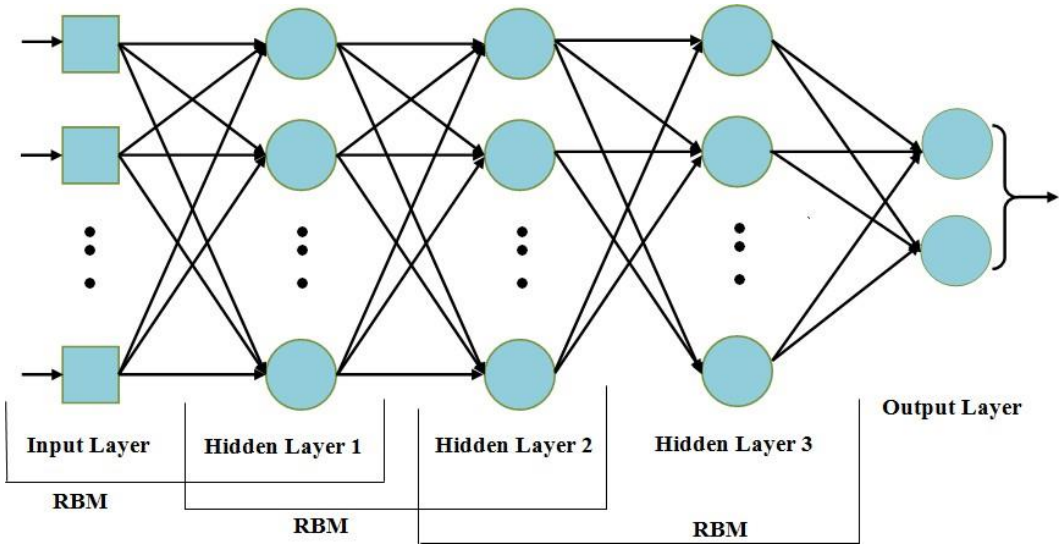

Figure 10: Working diagram of DBN

Evaluation

The outcomes of the proposed MR paradigm are investigated against the existing mechanisms. The work is executed in PHYTHON with Windows 7 OS, Intel i5/core processor, 4GM RAM and 3.20GHz CPU speed. The dataset used in the study is an openly available heart disease dataset and is accessed via Kaggle. Initially, the proposed KKM-based EBOA technique’s results for BD management in HDFS are analysed concerning accuracy, shown in Figure. 11. The system's accuracy is measured by considering a number of mappers and clusters in the MR framework.

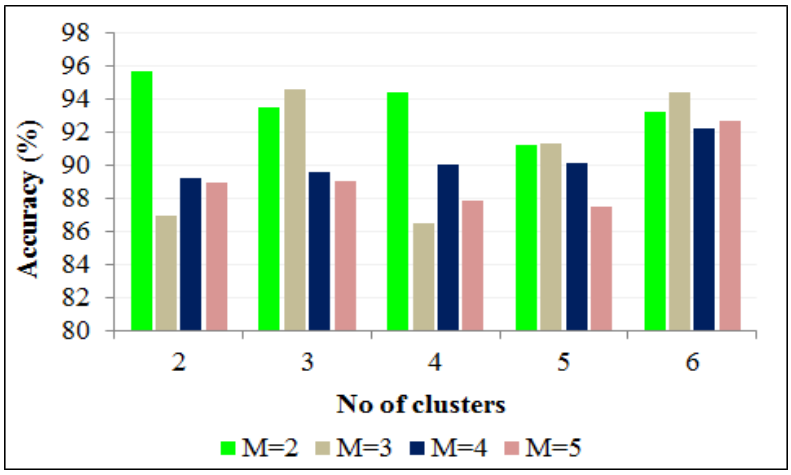

Figure 11: Accuracy Analysis

The method attains a clustering accuracy of 95.67% for the mapper size of 2 and the cluster size of 2. When the mapper size increased to 3,4 and 5, the clustering accuracy attained by our model was 86.89%, 89.18%, and 88.89%. Similarly, when we consider all mapper and cluster sizes, our model generates higher data management accuracy, showing its effectiveness in clustering the BD.

Next, the proposed EACWODBN is contrasted against the existing DBN, multi-layer perceptron (MLP), convolution neural network (CNN), and artificial neural network (ANN) models for BD classification regarding accuracy metric, which is plotted in Figure. 12. The accuracy of the models is plotted by fluctuating the simulations run from 1 to 5. Our model attains 97.22% accuracy for run 1, higher than the existing models. Similarly, our models show remarkable performance for BD classification for the remaining runs.

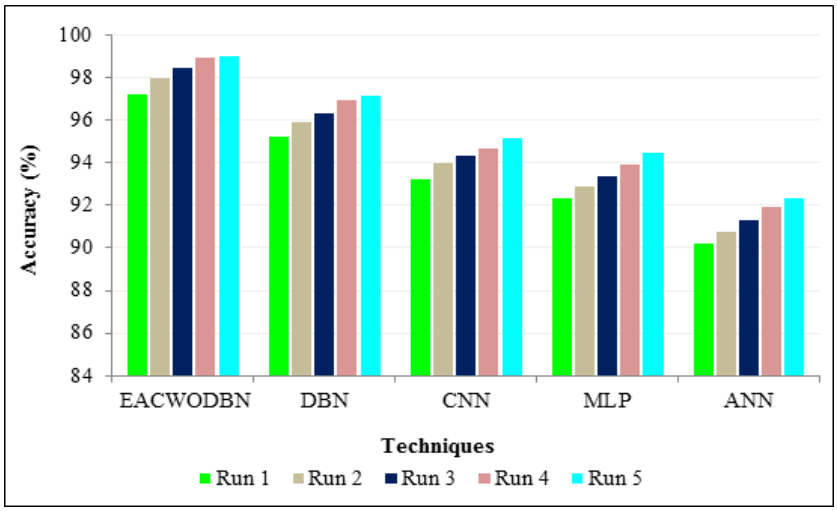

Figure 12: Accuracy analysis

Next, the performance of the security model called WHERSA is contrasted against the existing works regarding execution time, illustrated in Figure. 13. The results are plotted by changing the file sizes from 100 to 500. The proposed WHERSA attains the execution time of 44,354ms, 46,567ms, 49,497ms, 54,678ms, and 56,435ms for the 100, 200, 300, 400, and 500 MB data files, which are lower than the existing RSA, DSA, AES, data encryption standard (DES) and diffie hellman (DH) models.

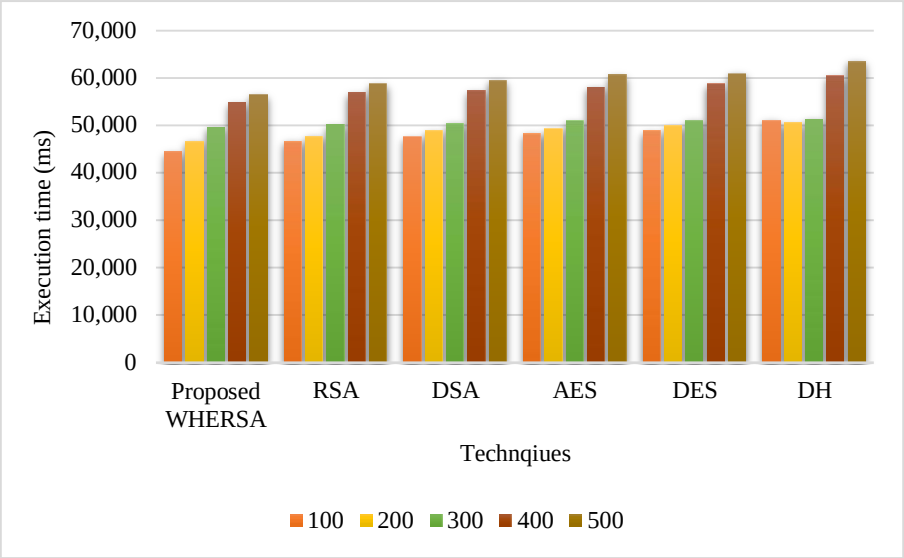

Figure 13: Evaluation based on execution time

It is observed from the outcomes that our encryption model outperforms other symmetric and asymmetric kind of encryption models. Compared to other models, ours is an asymmetric one and follows intricate mathematical procedures to carry out encryption as well as decryption. Thus, it is difficult to hack the information from the cloud storage. Also, our models use the smaller key to perform both encryptions together with decryption that delivers lower execution time. So, from the outcomes, our method optimally manages the BD in the cloud by offering a highly secure data transfer environment.

6. CONCLUSION AND FUTURE STUDIES

The study proposes a secure and optimal cluster-based data management scheme for patient monitoring in a cloud-enabled BD environment. The first research work develops an MFCM-COA-based optimal MR scheme for data management with 3FA and SHAXECC-based BD security in the cloud. The proposed MFCM-COA attains the highest accuracy of 96% for clustering the BD, and the encryption models prove its effectiveness with existing techniques.

The study develops another optimal and secure MR-based Big data management model with WHERSA in the second research work. An optimal DBN is also proposed to classify the medical Big data that aids in healthcare monitoring. The system reaches 95.67% higher accuracy beneath several cluster sizes and 99.01% for Big data classification. Finally, the WHERSA achieves higher security with lower execution time execution for secure cloud storage of the medical Big data. Thus, our scheme is determined to carry out secure Big data management more proficiently than the prevailing methods.

In future, we will study the challenges faced by the BD medical networks when operated in an Internet of things-based sensor environment for medical Big data management. They are vulnerable to several routing and security challenges due to their open nature. Also, the features like heterogeneity, constrained resources, uncontrolled environment, and scalability make the security difficulties even worse. So, we will prolong the work to propose an energy-aware intrusion detection technique to recognize the intrusions on the patients' healthcare data obtained via medical IoT sensors.

7. ORGANIZATION OF THESIS

Chapter 1: "Introduction" presents the notions of medical Big data, cloud computing, data management, HDFS, MR, research challenges, security issues, limitations of the existing techniques, and requirement of optimal clustering, encryption and access control mechanisms.

Chapter 2: *"Literature Review"* surveys the recent research works of various authors for Big data management, Big data security and Big data access control mechanisms in cloud environments.

Chapter 3: *"3FA and LFCOA based Optimal Clustering for Maintaining User Privacy in Cloud*" describes an optimal MR-based Big data management system in HDFS access control mechanisms for providing secure data communication over Big data based cloud platforms. The evaluation of the system is done by contrasting its outcomes with the existing techniques by considering several performance metrics.

Chapter 4: "*Secure Cloud Storage using SHAXECC"* describes the secure hashing-based encryption mechanism called SHAXECC for providing secure cloud storage of the medical big data. The evaluation of the system is done by contrasting its outcomes with the existing techniques by considering several performance metrics.

Chapter 5: "*Cluster-based Optimal MR model for BD Management and WHERSA based Secure Medical Storage*" describes the secure as well as optimal cluster-based Big data management system in HDFS and blockchain-based hashing mechanism for access control and encryption to deliver protected data communication in Big data -based cloud platforms. The optimal DBN is developed to perform Big data classification in the healthcare domain for patient monitoring. The evaluation of the system is done by contrasting its outcomes with the existing techniques by considering several performance metrics.

Chapter 6: "*Conclusions and Future Studies*" describes the conclusion of the thesis with future studies.

REFERENCES

1. Karim Abouelmehdi, Abderrahim Beni-Hessane, and Hayat Khaloufi, "Big healthcare data: preserving security and privacy", *Journal of big data*, **5** (1) (2018), 1-18.
2. Jinbo Xiong, Yuanyuan Zhang, Shaohua Tang, Ximeng Liu and Zhiqiang Yao, "Secure encrypted data with authorized deduplication in cloud", *IEEE Access*, **7** (2019), 75090-75104.
3. Shekha Chenthara, Khandakar Ahmed, Hua Wang and Frank Whittaker, "Security and privacy-preserving challenges of e-health solutions in cloud computing", *IEEE Access*, **7** (2019), 74361-74382.
4. H. Tanvir Sardar , and Zahid Ansari, "MapReduce-based fuzzy C-means algorithm for distributed document clustering", Journal of The Institution of Engineers (India): Series B, (2022), 1-12.
5. Chunqiong Wu, Bingwen Yan, Rongrui Yu, Baoqin Yu, Xiukao Zhou, Yanliang Yu, and Na Chen, "K-means clustering algorithm and its simulation based on distributed computing platform", Complexity, (2021), 1-10.
6. Noor S. Sagheer, and Suhad A. Yousif, "Canopy with k-means clustering algorithm for big data analytics", In AIP Conference Proceedings, **2334** (1) (2021).
7. Gourav Bathla, Himanshu Aggarwal, and Rinkle Rani, "A novel approach for clustering big data based on mapreduce", International Journal of Electrical and Computer Engineering (IJECE), **8** (3) (2018), 1711-1719.

8. Shaopeng Guan, Conghui Zhang, Yilin Wang, and Wenqing Liu, “Hadoop-based secure storage solution for big data in cloud computing environment”, Digital Communications and Networks (2023).

9. Gunasekar Thangarasu, Dominic P. D. D, Kayalvizhi Subramanian and Sajitha Smiley, “Biometric based signature authentication scheme for cloud healthcare data security”, **843** 557-565 (2019).

10. Fouad H. Awad, and Murtadha M. Hamad, “Improved k-means clustering algorithm for big data based on distributed smartphoneneural engine processor”, Electronics, **11** (6) (2022).

11. Usha Lawrancea and Jesu Vedha Nayahi Jesudhasan, “Privacy preserving parallel clustering-based anonymization for big data using mapreduce framework”, Applied Artificial Intelligence, **35** (15) (2021), 1587-1620.

12. Lakshmi Srinivasulu Dandugala, and Koneru Suvarna Vani, “A Weighted K-Means clustering Algorithm for Clustering Big Data based on MapReduce”, (2022).

13. Ramdas Vankdothu, Mohd Abdul Hameed, Raju Bhukya, and Gaurav Garg, “Entropy and sigmoid based K-means clustering and AGWO for effective big data handling”, Multimedia Tools and Application, **82**, (10) (2023), 15287-15304.

14. E Shanmugapriya and R Kavitha, “Efficient and secure privacy analysis for medical big data using TDES and MKSVM with access control in cloud”, Journal of Medical Systems, **43** (2019), 1-12.

15. Ila Chandrakar, and Vishwanath R. Hulipalled, “Improved technique for preserving privacy while mining real time big data”, International Journal of Communication Networks and Information Security, **14** (1) (2022), 86-92.

16. Siti Dhalila Mohd Satar, Mohamad Afendee Mohamed, Masnida Hussin, Zurina Mohd Hanapi and Siti Dhalila Mohd Satar, “Cloud-based secure healthcare framework by using enhanced ciphertext policy attribute-based encryption scheme”, (IJACSA) International Journal of Advanced Computer Science and Applications, **12** (1) (2021), 393-399.

17. Kashish A Shakil, Farhana J Zareen, Mansaf Alam and Suraiya Jabin, “BAMHealthCloud: A biometric authentication and data management system for

healthcare data in cloud", Journal of King Saud University – Computer and Information Sciences (In Press), (2017).

18. Sunday Oyinlola Ogundoyin and Ismaila Adeniyi Kamil, "PAASH: A privacy-preserving authentication and fine-grained access control of outsourced data for secure smart health in smart cities", Journal of Parallel and Distributed Computing (In Press), (2021).

19. Zhou, Y., & Varzaneh, M. G. Efficient and scalable patients clustering based on medical big data in cloud platform. *Journal of Cloud Computing*, **11** (1), (2022) 1-10.

20. Game, P. S., Vaze, V., & Emmanuel, M. Optimized Decision tree rules using divergence based grey wolf optimization for big data classification in health care. *Evolutionary Intelligence*, (2019) 1-17.

21. Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE access*, **8**, 107562-107582.

LIST OF PUBLICATIONS:

Journal of Publications – SCI

1. K Rajeshkumar, S Dhanasekaran, V Vasudevan, “A novel three-factor authentication and optimal mapreduce frameworks for secure medical big data transmission over the cloud with shaxecc”, Multimedia Tools and Applications, (2024), 1-29.
2. K Rajeshkumar, S Dhanasekaran, V Vasudevan, “Efficient and secure medical big data management system using optimal map-reduce framework and deep learning”, Multimedia Tools and Applications, (2023), 1-28.

Journal of Publications – SCOPUS

3. K Rajeshkumar, S Dhanasekaran, V Vasudevan, “Evaluation of risk analysis process in medical big data using Machine Learning”, International Conference on Computer Communication and Informatics, IEEE Access, (2021).
4. K Rajeshkumar, S Dhanasekaran, V Vasudevan, “Applications of machine learning algorithms for HDFS big data security”, International Conference on Computer Communication and Informatics, IEEE Access, (2022).

Journal of Publications- UGC

5. K Rajeshkumar, S Dhanasekaran, V Vasudevan, “Exploration of Big Data Security Framework using Machine Learning”, International Journal of Engineering Advanced and Technology,(2021).